

# A Review on Flight Delay Prediction

Alice Sternberg, Jorge Soares,  
Diego Carvalho, Eduardo Ogasawara \*

*CEFET/RJ*  
*Rio de Janeiro, Brazil*

March 20, 2017

## Abstract

Flight delays have a negative effect on airlines, airports and passengers. Their prediction is crucial during the decision-making process for all players of commercial aviation. Moreover, the development of accurate prediction models for flight delays became cumbersome due to the complexity of air transportation system, the amount of methods for prediction, and the deluge of data related to such system. In this context, this paper presents a thorough literature review of approaches used to build flight delay prediction models from the Data Science perspective. We propose a taxonomy and summarize the initiatives used to address the flight delay prediction problem, according to scope, data and computational methods, giving special attention to an increasing usage of machine learning methods. Besides, we also present a timeline of major works that depicts relationships between flight delay prediction problems and research trends to address them.

## 1 Introduction

Delay is one of the most remembered performance indicator of any transportation system. Notably, commercial aviation players understand delay as the period of time by which a flight is late or postponed. Thus, delay may be represented by the difference between scheduled and real times of departure or arrival of a flight [47]. Country regulator authorities have a multitude of indicators related to tolerance thresholds for flight delays. Indeed, flight delay is an important subject in the context of air transportation systems. In 2013, 36% of flights delayed by more than five minutes in Europe, 31.1% of flights delayed by more than 15 minutes in the United States and in Brazil, 16.3% of flights were canceled or suffered delays greater than 30 minutes [4, 16, 42]. This indicates how relevant this indicator is and how it affects no matter the scale of airline meshes.

Flight delays have negative impacts, particularly economic, for passengers, airlines and airports. Given the uncertainty of their occurrence, passengers usually plan to travel many hours before their appointments, increasing their trip costs, to ensure their arrival on time. On the other hand, airlines suffer penalties, fines and additional operation costs, such as crew and aircrafts retentions in airports [11]. Furthermore, from the sustainability point of view, delays may also cause environmental damage by increasing fuel consumption and gas emissions [34, 38, 40].

Delays also jeopardize airlines marketing strategies, since airlines rely on customers' loyalty to support their frequent-flyer programs and the consumer's choice is also affected by punctual performance [45]. There is a known relationship between levels

---

\*jsoares@cefet-rj.br, d.carvalho@ieee.org, eogasawara@ieee.org

of delays and fares, aircraft sizes, flight frequency and complains about airline service [9, 32, 53]. The estimation of flight delays can improve the tactical and operational decisions of airports and airlines managers and warn passengers, so they can rearrange their plans [27].

In an effort to better understand the entire flight ecosystems, huge volumes of data from commercial aviation are collected every moment and stored in databases. Submerged in this massive amount of data, analysts and data scientists are intensifying their computational and data management skills to extract useful information from each datum. In this context, the procedure of comprehending the domain, managing data and applying a model is known as Data Science, a trend in solving challenging problems related to Big Data.

Under this data deluge scenario, this paper contributes by presenting an analysis of the available literature on flight delay prediction from Data Science perspective. It seeks to summarize the most researched trends in this field, describing how this problem is addressed and comparing methods that have been used to build prediction models. This becomes more relevant as we observe an increasing presence of machine learning methods to model flight delays predictions. This analysis is conducted by establishing a flight delay research taxonomy, which organizes approaches according to type of problem, scope, data issues and computational methods. The paper also contributes by presenting a timeline of major works grouped by the type of flight delay prediction problem.

Besides this introduction, the rest of this paper is structured as follows. Section 2 introduces the flight delay scenario, describing a typical operation of a commercial flight, kinds of delays and their impacts. It also structures three different ways for treating the prediction problem. In Section 3, a taxonomic analysis of the prediction is presented, showing the most researched topics, the scope of application, data and methods that authors are using to predict flight delays. Section 4 discusses the main results based on a timeline of publications grouped by the types of problems and their intersections. Finally, Section 5 concludes our analysis presenting major highlights and trends about delay prediction problem.

## 2 The flight delay scenario

Commercial aviation is a complex distributed transportation system and it deals with expensive resources, demand fluctuations and an intricate origin-destination matrix that need orchestration to provide smooth and safety operations. Furthermore, individual passenger follows her itineraries while airlines plan various schedules for aircrafts, pilots and flight attendants [7]. Figure 1 illustrates a typical operation of a commercial flight. Stages can take place at terminal boundaries, airports, runways and airspace, being susceptible to different kinds of delays. Some examples include mechanical problems, weather conditions, ground delays, air traffic control, runway queues and capacity constraints [3, 21, 37].

This scheme is repeated several times throughout the day for each flight in the system. Pilots, flight attendants and aircrafts may have different schedules due to legal rests, duties, and maintenance plans for aircrafts. So, any disruption in the system can impact the subsequent flights of the same airline [2]. Moreover, disruptions may cause congestion at airspace or at other airports, creating queues and delaying some flights from other airlines [39, 51]. In this way, the prediction of flight delays is an important subject for airlines and airports.

The flight delay prediction problem can be treated by different points of view: (i) delay propagation, (ii) delay innovation and (iii) cancellation analysis. In delay propagation, one study how delay propagates through the network of the transportation

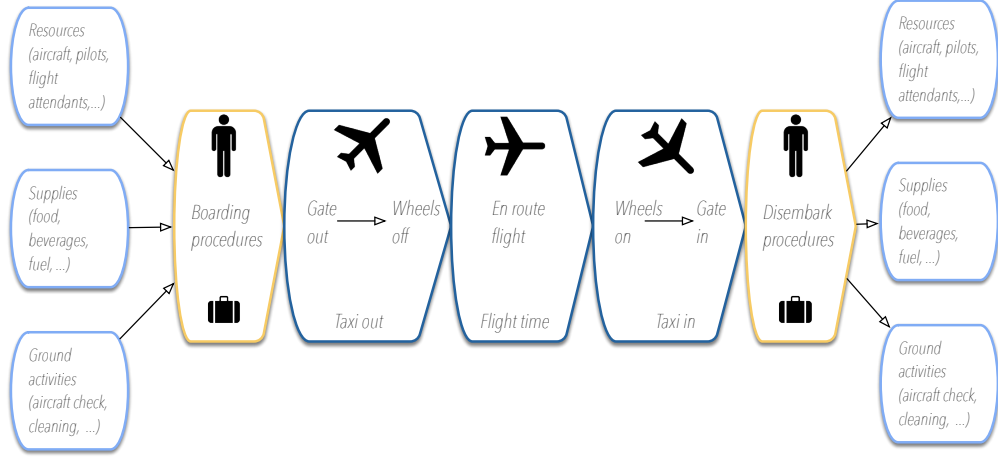


Figure 1: A typical operation of a commercial flight

system. On the other hand, considering that new problems may happen eventually, it is also important to predict new delays and understand their causes. Such occurrences is named as *delay innovation* problem. Finally, under specific situations, delays can lead to cancellations, forcing airlines and passengers to reschedule their itineraries. So, researches focused on cancellation analysis try to figure out which conditions lead to cancellations. Moreover, it explores the airlines decision making process for choosing the flights to be canceled.

### 3 Taxonomy

The main problems related to flight delay prediction are identified and organized in a taxonomy. It includes scopes, models, and ways of handling flight delay prediction problem. It considers flight domain features, such as *problem* and *scope*, and Data Science perspective, such as *data* and *methods*. Figure 2 depicts the entire taxonomy while next subsections describe each component of the taxonomy and related researches.

Regarding the available literature on flight delay prediction, we have conducted a systematic mapping study. The search expression string  $(\text{"airport delay"} \vee \text{"flight delay"}) \wedge (\text{"predict"} \vee \text{"forecast"} \vee \text{"propagate"})$  was used to query Science Direct and IEEE databases on January 2015. Query result brought 448 references, from which 28 were selected to this analysis due to their relevance and direct link with the flight delay prediction problem. Additionally, five works were added using backward snowballing search. Thus, the presented taxonomy is based on 33 publications, and the following subsections present the devised taxonomy from the final query result.

#### 3.1 Problem

Problem is the core feature in domain taxonomy. As seen in Section 2, there are three major concerns regarding the flight delay prediction problem: delay propagation, delay innovation and cancellation analysis. Depending on the emphasis of the research, authors select one of these lines to develop their models.

##### 3.1.1 Delay propagation

In delay propagation the main objective is to understand how delay propagates through airlines and airports based on the assumption that a initial delay have already occurred

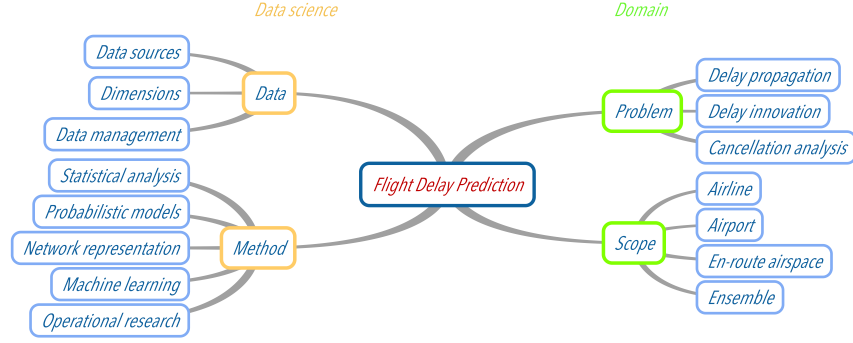


Figure 2: Taxonomy of the flight delay prediction problem

in the transportation system. A particular scenario occurs when delays are propagated to other flights of the same airline as chain reactions [2, 8, 10, 48]. Under this scenarios, it is important to measure how stable and reliable airlines can be in order to recovery from delay propagation [15, 49]. In addition, delay may continue to propagate due to scheduling of critical resources or retentions in other airports [19].

When scheduled time for take-off or landing is not fulfilled, flights need new slots that may be unavailable. In this scenario, it is important to understand the effects that a root delay in a flight may produce to both departure and arrival airports [20, 35, 51]. Such phenomenon may increase the number of flights at some period, generating capacity problems and queues.

### 3.1.2 Delay Innovation

Considering that new delays may happen eventually, these delay innovations impair the performance of transportation network. Researches create prediction models to tackle delay innovations, predicting when and where a delay will occur and what are its reasons and sources. This includes models that effectively seek to estimate the number of minutes, probability or level of delay for a specific flight, airline or airport.

A relevant number of works focused on predicting and estimating delay duration [26, 36]. Some approaches considered probabilistic models and innovation distribution [29, 43], whereas others consider conditions for the occurrence of delay innovation, such as passenger demand, fares, flight frequency, aircraft size, and taxi-out time [5, 22, 52].

### 3.1.3 Cancellation analysis

Particular circumstances, such as weather conditions, acts of God, aircraft problems, etc may lead airlines to cancel flights. Besides, airlines may directly cancel a flight, when factors like seat occupancy or cost savings are taking into consideration [50].

## 3.2 Scope

Delays can be induced by different sources and affect airports, airlines, *en route* airspace or an ensemble of them. For analysis purposes, one may assume a simplified system where only one of these actors or any combination of them is considered. It is should be noted that any scope of application can be combined to any type of problem mentioned in Section 3.1.

Some researches focused on airports to predict delays for all departs considered all airlines and *en route* airspace indifferently [36, 39]. Airports are also the focus when the objective is to investigate their efficiency based on delays of all airlines [24, 33, 35]. On

the other hand, only airlines are considered when comparing the performance of two airlines under the same conditions [3].

An ensemble of airport and *en route* airspace were studied to understand the relationship between congestion and delays [21, 22, 28]. Others considered airports and airlines as well to evaluate capacity problems and airlines decisions [43]. There are many possibilities to ensemble scopes. This becomes important when studying the dynamics of air transportation systems, particularly when targeting delay innovation.

### 3.3 Data

Three basic questions about data are: Where to find flight data? Which attributes should be considered? Is it possible to handle each datum in order to obtain better results? To answer these questions, the data problem is divided into (i) data sources: which presents the main sources of data used to develop the prediction models, (ii) dimensions: which displays data model for datasets linked to the flight delay prediction, and (iii) data management: which describes database concepts and preprocessing techniques applied to empower prediction models and improve their performance.

#### 3.3.1 Data Sources

Datasets from air transportation system are mainly available from governmental agencies, regulatory authorities, airlines, airports, and service providers. Table 1 displays the sources and regions considered by the publications analyzed in this paper. Governmental agencies usually provide public access to their databases with different granularity. It is noticed that data from The United States Department of Transportation [14], especially through The Federal Aviation Administration [18] and The Bureau of Transportation Statistics databases [12] are widely used to obtain information about flights. The Eurocontrol [17] database is provided by an intergovernmental organization in Europe. This dataset is also used intensively in flight delay studies [37].

Other related datasets, such as weather, may be obtained from governmental databases. This includes, for example, The National Oceanic and Atmospheric Administration of the United States [30]. Since airlines and airports do not share their databases with the entire community, they are often used by collaborators of those institutions. Additionally, some dataset are offered by service providers, such as The Weather Company [44]. In fact, authors may use more than one source to develop their models. Datasets from United States Department of Transportation [14], National Oceanic and Atmospheric Administration [30] and Weather Company [44] are commonly used to build delay prediction models.

Additionally, some researches [25, 52] create synthetic datasets to evaluate their models instead of using real data. For example, Zou et al. [52] developed a market scenario, considering airport capacity, links, frequency and characteristics of flights and passenger demand.

#### 3.3.2 Dimensions

Considering the main public datasets and the papers analyzed, we have organized them main commonly attributes used into seven classes depicted in the data model of Figure 3. They abstract the main input attributes for delay prediction models. Beyond scheduled and actual times of departure and arrival, several attributes may be considered depending on the focus of research.

Spatial dimension is related to the positions taken by the aircraft, such as departure and arrival airports, their cities, regions and countries [20, 36]. Temporal dimension is often used to capture seasonality or periodic patterns of data. These elements contain both date (season, month, and day of the week) and time (period of the day or time

Table 1: Sources of real data about the air transportation system per region

Source / Region	United States	Europe	Asia
Governmental Agencies	Boswell et al. [10]		
	Pathomsiri et al. [33]		
	Mueller et al. [29]		
	Tu et al. [43]		
	Wang et al. [46]		
	Balakrishna et al. [5]		
	Xu et al. [51]		
	Kim et al. [24]		
	Abdel-Aty et al. [1]	Feighan et al. [37]	
	Pyrgiotis et al. [35]		
	Hunter et al. [21]		
	Xiong et al. [50]		
	Hunter et al. [22]		
	Hao et al. [20]		
	Balakrishna et al. [6]		
	Khanmohammadi et al. [23]		
	Morrison et al. [28]		
	Rebollo et al. [36]		
Airlines	Beatty et al. [8]		
	AhmadBeygi et al. [3]	Wu [49]	Wong et al. [48]
	Abdelghany et al. [2]	Duck et al. [15]	
Airports		Soomer [41]	Lu [26]
Service providers	Wieland [47]		
	Hunter et al. [21]	Feighan et al. [37]	
	Schaefer et al. [39]		
	Hunter et al. [22]		
	Hansen [19]		

of the day) characteristics [1, 29, 43]. Weather dimension expresses external and environmental conditions in a certain moment. It may represent specific characteristics, such as ceiling and visibility [37] that defines, for example, if take-off or landing is going to happen under visual or instrumental conditions. Additionally, *en route* airspace weather situation (known as convective weather) and airport weather situation (known as surface weather) contain several momentaneous parameters [21, 22].

Planning describes what airlines, airports and air traffic controllers intend to do with critical resources involved in their operations. This dimension includes: (i) airline schedules, (ii) airport schedules and (iii) flight plans. Arline schedules define all origin and destination points, their frequency and sequence, and aircrafts and crew allocations for each flight [3, 8, 10, 15, 49]. Airport schedules indicate the time each flight takes-off and lands, while flight plans indicate all *en route* parameters, such as distance, route, speed, and high [19].

Features represent characteristics of airlines, airports or aircrafts. Airlines status may indicate if an airline is a major or an affiliate one or if it is a traditional hub-and-spoke or a low cost point-to-point. Aircrafts characteristics indicate their size, their number of seats and occupancy, which may be a constraint to some operations because they affect market decisions. Finally, airport infrastructure may represent the number

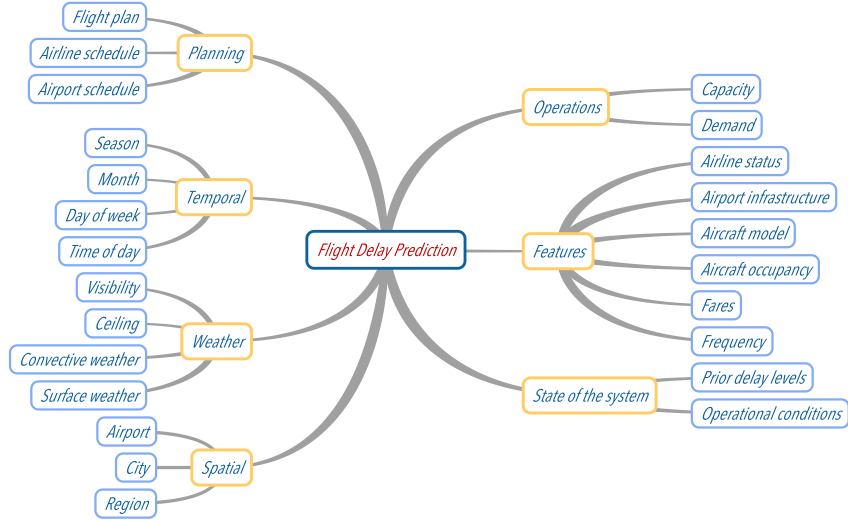


Figure 3: Data model of the flight delay prediction

of runways, gates and service providers in an airport facility [33, 41, 50].

The state of the system indicates in which conditions airlines, airports or *en route* airspace are operating in a specific moment. Some examples correspond to prior levels of delay or airports closures [25]. The information about the state of the system is useful to predict its behavior. Finally, operations are related to capacity and demand of airports and *en route* airspace. When demand exceeds capacity, a congestion scenario is formed, which enables occurrence of delays [28].

### 3.3.3 Data Management

Since the use of databases to store huge amount of data have been increasing over the last years, data management techniques are becoming more and more crucial to provide a convenient and efficient query processing. Data management tasks contemplate design of database structure to enable data integration from different sources, elimination of inconsistencies, and data transformation. The development of a data warehouse supported by online analytical processing (OLAP) and data management techniques may be useful for this purpose. As mentioned in Section 3.3.1, multiple sources of data may be used. Thus, the usage of data warehouses combined with Extract, Transform and Load (ETL) procedures are commonly used to link the datasets of different sources.

There are many data management preprocessing procedures that can be applied to flight delay prediction datasets. They include data cleaning, feature selection, data transformation, and clustering. One of the main tasks in data cleaning is outlier removal. Extreme conditions may result in outliers that are not interesting if one is concerned about regular operations [43]. Feature selection is the process of identifying attributes that are less correlated. Correlated and irrelevant attributes may provide model over-fitting or decrease prediction performance [48]. These preprocessing procedures are important, since the better the preprocessing is conducted on input data, the better the prediction models may be developed from it.

Data transformation is also a important activity to empower prediction models. Some examples of transformations includes normalization and discretization. Normalization reduces the range of possible values to a particular interval, such as -1 to 1 or 0 to 1. This is important to give equal strength for different variables and let machine learning methods identify which are the most relevant ones [25, 31]. Discretization con-

sists of replacing numerical values by representative labels. This includes the transform of time periods into bins of fixed time [5, 24], binning of values to cope with limitations in computational packages [10, 51] or to better train prediction models [8], specially when using machine learning models.

Clustering means grouping elements of the dataset in a way that similar observations stay together in the same group and dissimilar items stay in different groups. Many researches compute clustering techniques, such as k-means or agglomerative hierarchical clustering, to support preliminary steps for further prediction models [26, 36].

### 3.4 Method

The flight delay prediction problem may be modeled by many methods, depending on the objectives of the researches. Methods were divided in five groups, according to Figure 4. The numbers next to each category represent the amount of related papers.

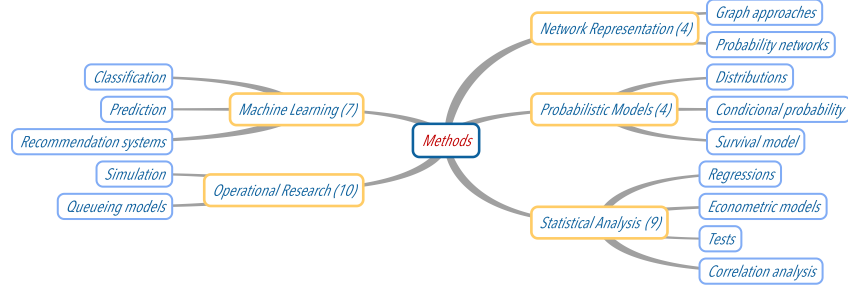


Figure 4: Categories of methods used to model the flight delay prediction

#### 3.4.1 Statistical Analysis

Statistical analysis usually encompasses the usage of regression models, correlation analysis, econometric models, parametric tests, non-parametric tests, and multivariate analysis (MVA). When it comes to regression models, both delay multiplier and recursive models can help airlines to understand delay propagation effects through the network and to estimate the costs of delays [8, 46].

Many econometric models are also build to evaluate the efficiency flight systems, such as the analysis of the investments done by a governmental agency [28] or to evaluate the equilibrium point considering the relationship between delays and passenger demand, fares, frequency and size of the aircrafts [52]. Xiong et al. [50] built an econometric model based on pre-existing delays, potential delay savings, distance, characteristics of the destination airport and airline, frequency, aircraft size, occupancy rate and fare to understand which reasons lead airlines to cancel their flights. Finally, Hao et al. [20] built a model to quantify how delays originated at New York are propagated to other airports.

Some works focus on statistical inference. Pathomsiri et al. [33] used a non-parametric function to evaluate the efficiency of airports of the United States in terms of delays. Reynolds et al. [37] computed the correlation between levels of delays and capacities of the European airports. They also suggested different approaches to deal with the congestion problem, describing their advantages and disadvantages. Finally, Abdel-Aty et al. [1] calculated daily average of delays to detect correlations to understand the main causes of delays at Orlando International Airport.



### 3.4.2 Probabilistic Models

Probabilistic Models encompass analysis tools that estimate the probability of an event based on historical data. Tu et al. [43] developed a probabilistic model based on expectation-maximization combined with genetic algorithms to estimate the distribution of departure delay at Denver International Airport.

Boswell et al. [10] expressed delay classes by a probabilistic mass function and used a transition matrix to verify delay propagation to subsequent flights. They made a cancellation analysis computing the conditional probability to cancel a flight given that its preceding flight was delayed. Mueller et al. [29] modeled departure, *en route* and arrival delays using density functions. The authors verified that Normal distribution fitted better to departure delays, while *en route* and arrival delays were better described by Poisson distribution. Concerned about the total duration of a root delay, Wong et al. [48] studied delay propagation through a survival model.

### 3.4.3 Network Representation

Network representation encompasses the study of flight systems according to a graph theory. Abdelghany et al. [2] built direct acyclic graphs to model the schedule of an airline (including flight times and resources availability) to detect disruptions and their impacts on the rest of the network. They used the classical shortest path algorithm to evaluate propagation effects. Ahmadbeygi et al. [3] built propagation trees to compare two different airlines, one operating in a conventional hub-and-spoke scheme and the other in a low-cost point-to-point system. Finally, Xu et al. [51] built a Bayesian network to model delay propagation at three airports in the United States.

### 3.4.4 Operational Research

Operational Research includes advanced analytical methods (such as optimization, simulations, and queue theory) to help key-players make better decisions. Simulations may analyze airport capacity data, considering departure and arrival delays under different weather conditions [21, 22, 39]. They may also evaluate the cost of each delayed flight of an airline schedule [41]. Moreover, simulations through queuing models were applied by Wieland [47] to predict delay innovation, by Kim and Hansen [24] to study the effects of capacity and demand on delay levels at the airports of New York area, and by Pyrgiotis et al. [35] to study delay propagation between some airports.

Other simulations were done to analyze delay propagation concerning schedule stability [15] and reliability [49]. Through simulations, different scenarios were commonly explored, such as reliability or flexibility of airports under external conditions. Hansen et al. [19] considered the congestion problem and designed a simple deterministic queuing model to analyze propagation effects for subsequent flights of an airline and at Los Angeles International Airport.

### 3.4.5 Machine Learning

Machine learning is the research that explores the development of algorithms that can learn from data and provide predictions from it. Researches that study flight systems are increasing the usage of machine learning methods. The methods commonly used includes k-Nearest Neighbor, neural networks, SVM, fuzzy logic, and random forests. They were mainly used for classification and prediction.

Rebollo et al. [36] applied random forests to predict delay innovation. They compared their approach with regression models to predict delay innovation in airports of the United States considering time horizons of 2, 4, 6 and 24 hours. Their test errors grew as the forecast horizon increased. Also in delay innovation, Lu et al. [26] compared

the performances of Naïve Bayes, decision tree and neural network to predict delays in large datasets. They observed that decision tree had the best performance presenting a prediction confidence of 80%.

Chen et al. [13] built a model based on a fuzzy support vector machine with weighted margin for predicting departure and arrival delays. The authors defined five grades of delay and concluded that the performance of an ensemble fuzzy support vector machine with weighted margin was more accurate than a standalone support vector machine. Khanmohammadi et al. [23] created an adaptive network based on fuzzy inference system to predict delay innovations. The predictions were used as an input for a fuzzy decision making method to sequence arrivals at JFK International Airport in New York.

Balakrishna et al. [5, 6] used a reinforcement learning algorithm to predict taxi-out delays. The problem was modeled through a Markov decision process and solved by a machine learning algorithm. When running their model 15 minutes before the scheduled time of departure, authors achieved good performances at JFK International Airport in New York and at Tampa Bay International Airport.

Lu et al. [25] built a recommendation system to forecast delays at some airports due to propagation effects. Prediction was based on the k-Nearest Neighbor algorithm and used historical data to recognize similar situations in the past. The authors noticed fast response time and easy logical comprehension as the main advantages of their method.

## 4 Results and discussion

Since flight delays cause economic consequences to passengers and airlines, recognizing them through prediction may improve marketing decisions. Due to that, several forecast models have been built over the last twenty years. These models have sought to understand how delays propagate through the network of flights or airports, to predict delay innovations in the system or to comprehend the cancellation process. Beyond these three points of view for treating the flight delay prediction problem, models could also differ by their scope of application, data issues and methods.

The number of papers has increased in the late 2000s, since 67% of the researches had been published between 2007 and 2014. Regarding only the papers considered in this analysis, Figure 5 displays the complete timeline of publications, showing authors, regions of the datasets and categories of methods applied to flight delay models.

Pondering the way for tackling the delay problem, it was seen a balance between the number of papers that consider delay propagation and delay innovation, while few researches considered the cancellation analysis. These three lines may be combined with any scope of application. For example, both delay propagation and delay innovation models were built based on an ensemble of airports and *en route* airspace elements.

According to data perspective, we divided our analysis into three parts: data sources, dimensions and data management. From our review analysis, the adoption of data sources depends mostly of the country or region where the study has been taken place, mostly conducted by researches of that particular region. For example, in China, most researches were based on airport data, while in the United States the main source was The United States Department of Transportation [14].

Dimensions were not directly related to type of problem, but to the scope of application. This characteristic is notable in this case. Attributes such as weather, capacity and demand were characteristics of airport or *en route* airspace scopes. On the other hand, airlines schedules indicated scopes that considered airlines elements. It was also observed several ensembles of different dimensions, showing that prediction models may be improved through the selection of different attributes.

Data management was not specific of any problem or scope of application and its use is steadily growing. In fact, it is constantly present in most of the machine learning

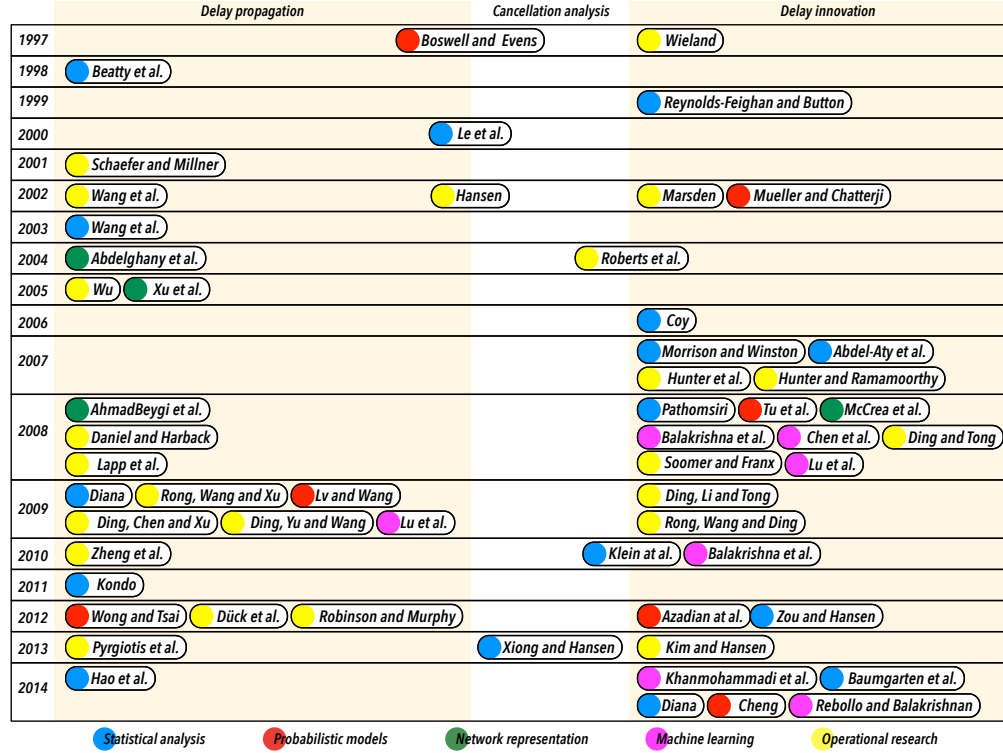


Figure 5: Time line of flight delay prediction publications

models adopted, especially through data transformation. Most of the probabilistic models also considered outlier removal and data transformations techniques. A small percentage of the statistical analysis, network representation and operational research methods applied general data management techniques as well.

Regarding the methods used to develop the forecast models, statistical analysis and operational research were the most applied in the past. These approaches were well spread between the three ways for treating the prediction problem. This same balance was also verified for probabilistic models. On the other hand, network representation was mostly employed for delay propagation.

It is worth mentioning that machine learning approaches experienced a notable growth in the late 2000s, especially in delay innovation. In fact, both machine learning and data management are positively correlated. The more machine learning is used, the more data management is required.

## 5 Conclusion

Flight delays are an important subject in the literature due to their economical and environmental impacts. They may increase ticket costs to customers and operational costs to airlines. Apart from outcomes directly related to passengers, delay prediction is crucial during the decision-making process for every player in the air transportation system.

In this context, researchers created flight delay models for delay prediction over the last years and this work contributes with an analyzes of these models from a Data Science perspective. We developed a taxonomy scheme and classified models in respect of detailed components.

Mainly, the taxonomy includes domain and Data Science branches. The former branch categorizes the problem (flight delay prediction) and the scope. The last branch groups methods and data handling. It was observed that the flight delay prediction is classified in three main categories, such as delay propagation, delay innovation, and cancellation analysis. Besides, the scope determines one of the three specific extents: airline, airport, en-route airspace or an ensemble of them.

Additionally, considering Data Science branch, we aimed at the datum, by categorizing data sources, dimensions that can be used in the models, and data management techniques to preprocess data and improve prediction models efficiency. We also studied and divided the main methods into five categories: statistical analysis, probabilistic models, network representation, operations research, and machine learning. Those categories have been grouped as their use on specific forecast models for flight delays.

Besides the taxonomic scheme, we also presented a timeline with all articles to spot trends and relationships involving the main elements in the taxonomy. In the light of the domain-problem classification, this timeline showed a dominance of delay propagation and delay innovation over cancelation analysis. Researchers used to focus at statistical analysis and operational research approaches in the past. However, as the data volume grows, we noticed the usage of machine learning and data management is increasing significantly. This clearly characterizes a Data Science trend.

Researchers from airlines, airports and academia will require a combination of skills of both domain specialists and data scientists to enable knowledge discovery from flight Big Data.

## Acknowledgments

The authors thank to CNPq, CAPES, and FAPERJ for partially funding this research.

## References

- [1] M. Abdel-Aty, C. Lee, Y. Bai, X. Li, and M. Michalak. Detecting periodic patterns of arrival delay. *Journal of Air Transport Management*, 13(6):355 – 361, 2007.
- [2] K. F. Abdelghany, S. S. Shah, S. Raina, and A. F. Abdelghany. A model for projecting flight delays during irregular operation conditions. *Journal of Air Transport Management*, 10(6):385 – 394, 2004.
- [3] S. AhmadBeygi, A. Cohn, Y. Guan, and P. Belobaba. Analysis of the potential for delay propagation in passenger airline networks. *Journal of Air Transport Management*, 14(5):221 – 236, 2008.
- [4] ANAC. Agncia Nacional de Aviao Civil. Technical report, <http://www.anac.gov.br/>, 2015.
- [5] P. Balakrishna, R. Ganesan, and L. Sherry. Accuracy of reinforcement learning algorithms for predicting aircraft taxi-out times: A case-study of Tampa Bay departures. *Transportation Research Part C: Emerging Technologies*, 18(6):950 – 962, 2010.
- [6] P. Balakrishna, R. Ganesan, L. Sherry, and B. Levy. Estimating Taxi-out times with a reinforcement learning algorithm. In *Digital Avionics Systems Conference, 2008. DASC 2008. IEEE/AIAA 27th*, pages 3.D.3–1–3.D.3–12, Oct. 2008.
- [7] C. Barnhart and G. Laporte. *Transportation*, volume 14. Elsevier, 2007.

- [8] R. Beatty, R. Hsu, L. Berry, and J. Rome. Preliminary evaluation of flight delay propagation through an airline schedule. *2nd USA/Europe Air Traffic Management R&D Seminar*, 7(4):259–270, 1998.
- [9] D. Bhadra. You (expect to) get what you pay for: A system approach to delay, fare, and complaints. *Transportation Research Part A: Policy and Practice*, 43(9-10):829 – 843, 2009.
- [10] S. B. Boswell and J. E. Evans. *Analysis of downstream impacts of air traffic delay*. Lincoln Laboratory, Massachusetts Institute of Technology, 1997.
- [11] R. Britto, M. Dresner, and A. Voltes. The impact of flight delays on passenger demand and societal welfare. *Transportation Research Part E: Logistics and Transportation Review*, 48(2):460 – 469, 2012.
- [12] BTS. The Bureau of Transportation Statistics databases. Technical report, <http://www.rita.dot.gov/bts/home>, 2015.
- [13] H. Chen, J. Wang, and X. Yan. A Fuzzy Support Vector Machine with Weighted Margin for Flight Delay Early Warning. In *Fuzzy Systems and Knowledge Discovery, 2008. FSKD '08. Fifth International Conference on*, volume 3, pages 331–335, Oct. 2008.
- [14] DOT. The United States Department of Transportation. Technical report, <http://www.dot.gov/>, 2015.
- [15] V. Dck, L. Ionescub, N. Kliwerb, and L. Suhla. Increasing stability of crew and aircraft schedules. *Transportation Research Part C: Emerging Technologies*, 20(1):47 – 61, 2012.
- [16] EUROCONTROL. CODA Digest - Delays to Air Transport in Europe 2013. Technical report, <http://www.eurocontrol.int/publications/coda-digest-annual-2013>, 2014.
- [17] EUROCONTROL. European Organisation for the Safety of Air Navigation. Technical report, <https://www.eurocontrol.int/>, 2015.
- [18] FAA. Federal Aviation Administration. Technical report, <http://www.faa.gov/>, 2015.
- [19] M. Hansen. Micro-level analysis of airport delay externalities using deterministic queuing models: a case study. *Journal of Air Transport Management*, 8(2):73 – 87, 2002.
- [20] L. Hao, M. Hansen, Y. Zhang, and J. Post. New York, New York: Two ways of estimating the delay impact of New York airports. *Transportation Research Part E: Logistics and Transportation Review*, 70(0):245 – 260, 2014.
- [21] G. Hunter, B. Boisvert, and K. Ramamoorthy. Advanced national airspace traffic flow management simulation experiments and vldation. In *Simulation Conference, 2007 Winter*, pages 1261–1267, Dec. 2007.
- [22] G. Hunter and K. Ramamoorthy. Evaluation of the national airspace system aggregate performance sensitivity. In *Digital Avionics Systems Conference, 2007. DASC '07. IEEE/AIAA 26th*, pages 1.E.1–1–1.E.1–13, Oct. 2007.
- [23] S. Khanmohammadi, C.-A. Chou, H. Lewis, and D. Elias. A systems approach for scheduling aircraft landings in JFK airport. In *Fuzzy Systems (FUZZ-IEEE), 2014 IEEE International Conference on*, pages 1578–1585, July 2014.

- [24] A. Kim and M. Hansen. Deconstructing delay: A non-parametric approach to analyzing delay changes in single server queuing systems. *Transportation Research Part B: Methodological*, 58(0):119 – 133, 2013.
- [25] Z. Lu, J. Wang, and T. Xu. A new method for flight delays forecast based on the recommendation system. In *Computing, Communication, Control, and Management, 2009. CCCM 2009. ISECS International Colloquium on*, volume 1, pages 46–49, Aug. 2009.
- [26] Z. Lu, J. Wang, and G. Zheng. A New Method to Alarm Large Scale of Flights Delay Based on Machine Learning. In *Knowledge Acquisition and Modeling, 2008. KAM '08. International Symposium on*, pages 589–592, Dec. 2008.
- [27] X. Lv and H. Wang. Flight Delay Alarming Analysis for an Airport Based on Markov. In *Education Technology and Computer Science, 2009. ETCS '09. First International Workshop on*, volume 1, pages 685–688, Mar. 2009.
- [28] S. A. Morrison and C. Winston. The effect of {FAA} expenditures on air travel delays. *Journal of Urban Economics*, 63(2):669 – 678, 2008.
- [29] E. R. Mueller and G. B. Chatterji. Analysis of aircraft arrival and departure delay characteristics. In *AIAA aircraft technology, integration and operations (ATIO) conference*, 2002.
- [30] NOAA. National Oceanic and Atmospheric Administration. Technical report, <http://www.noaa.gov/>, 2015.
- [31] E. Ogasawara, L. Murta, G. Zimbrao, and M. Mattoso. Neural networks cartridges for data mining on time series. In *International Joint Conference on Neural Networks, 2009. IJCNN 2009*, pages 2302–2309, June 2009.
- [32] V. Pai. On the factors that affect airline flight frequency and aircraft size. *Journal of Air Transport Management*, 16(4):169 – 177, 2010.
- [33] S. Pathomsiri, A. Haghani, M. Dresner, and R. J. Windle. Impact of undesirable outputs on the productivity of {US} airports. *Transportation Research Part E: Logistics and Transportation Review*, 44(2):235 – 259, 2008.
- [34] T. Pejovic, R. B. Noland, V. Williams, and R. Toumi. A tentative analysis of the impacts of an airport closure. *Journal of Air Transport Management*, 15(5):241 – 248, 2009.
- [35] N. Pyrgiotis, K. M. Malone, and A. Odoni. Modelling delay propagation within an airport network. *Transportation Research Part C: Emerging Technologies*, 27(0):60 – 75, 2013.
- [36] J. J. Rebollo and H. Balakrishnan. Characterization and prediction of air traffic delays. *Transportation Research Part C: Emerging Technologies*, 44(0):231 – 241, 2014.
- [37] A. J. Reynolds-Feighan and K. J. Button. An assessment of the capacity and congestion levels at European airports. *Journal of Air Transport Management*, 5(3):113 – 134, 1999.
- [38] M. S. Ryerson, M. Hansen, and J. Bonn. Time to burn: Flight delay, terminal efficiency, and fuel consumption in the National Airspace System. *Transportation Research Part A: Policy and Practice*, 69(0):286 – 298, 2014.

- [39] L. Schaefer and D. Millner. Flight delay propagation analysis with the Detailed Policy Assessment Tool. In *Systems, Man, and Cybernetics, 2001 IEEE International Conference on*, volume 2, pages 1299–1303 vol.2, 2001.
- [40] T. K. Simi and O. Babi. Airport traffic complexity and environment efficiency metrics for evaluation of {ATM} measures. *Journal of Air Transport Management*, 42(0):260 – 271, 2015.
- [41] M. J. Soomer and G. J. Franx. Scheduling aircraft landings using airlines’ preferences. *European Journal of Operational Research*, 190(1):277 – 291, 2008.
- [42] The Unites States Department of Transportation. Air Travel Consumer Report 2013. Technical report, <http://www.dot.gov/airconsumer/air-travel-consumer-reports>, 2014.
- [43] Y. Tu, M. O. Ball, and W. S. Jank. Estimating flight departure delay distributionsa statistical approach with long-term trend and short-term pattern. *Journal of the American Statistical Association*, 103(481):112–125, 2008.
- [44] TWC. The Weather Company. Technical report, <http://www.theweathercompany.com/>, 2015.
- [45] I. Vlachos and Z. Lin. Drivers of airline loyalty: Evidence from the business travelers in China. *Transportation Research Part E: Logistics and Transportation Review*, 71(0):1 – 17, 2014.
- [46] P. Wang, L. Schaefer, and L. Wojcik. Flight connections and their impacts on delay propagation. In *Digital Avionics Systems Conference, 2003. DASC '03. The 22nd*, volume 1, pages 5.B.4–5.1–9 vol.1, Oct. 2003.
- [47] F. Wieland. Limits to growth: results from the detailed policy assessment tool [air traffic congestion]. In *Digital Avionics Systems Conference, 1997. 16th DASC., AIAA/IEEE*, volume 2, pages 9.2–1–9.2–8 vol.2, Oct. 1997.
- [48] J.-T. Wong and S.-C. Tsai. A survival model for flight delay propagation. *Journal of Air Transport Management*, 23(0):5 – 11, 2012.
- [49] C.-L. Wu. Inherent delays and operational reliability of airline schedules. *Journal of Air Transport Management*, 11(4):273 – 282, 2005.
- [50] J. Xiong and M. Hansen. Modelling airline flight cancellation decisions. *Transportation Research Part E: Logistics and Transportation Review*, 56(0):64 – 80, 2013.
- [51] N. Xu, G. Donohue, K. B. Laskey, and C.-H. Chen. Estimation of delay propagation in the national aviation system using Bayesian networks. In *6th USA/Europe Air Traffic Management Research and Development Seminar*. Citeseer, 2005.
- [52] B. Zou and M. Hansen. Flight delays, capacity investment and social welfare under air transport supply-demand equilibrium. *Transportation Research Part A: Policy and Practice*, 46(6):965 – 980, 2012.
- [53] B. Zou and M. Hansen. Flight delay impact on airfare and flight frequency: A comprehensive assessment. *Transportation Research Part E: Logistics and Transportation Review*, 69(0):54 – 74, 2014.